Learning to Rank Trees in a Heterogeneous Graph with Applications in Investigative Journalism

Oana Balalau[1], Ioana Manolescu[1], and Fabian Suchanek[2]

[1]Inria Saclay, firstname.lastname@inria.fr
[2]Télécom Paris, firstname@lastname.name

Keywords: machine learning, learning to rank, graph embeddings, crowdsourcing
Team: CEDAR team, Inria SIF and LIX (CNRS UMR 7161 and Ecole polytechnique); DIG team, Télécom Paris.

Background: In this project, we are interested in mining useful information from large datasets, in order to provide support for investigative journalism. Real-world events such as elections, public demonstrations, disclosures of illegal or surprising activities, etc. are mirrored in new data items being created and added to the global corpus of available information. Making sense of this wealth of data by finding important connections between entities of interest could facilitate the work of journalists. The team has an on-going collaboration with fact-checking journalists from Le Monde.

In collaboration with Le Monde, the team CEDAR has collected public data about French politicians and has organized this information as a heterogeneous graph, where nodes represent entities, i.e. person, location or organization, and edges represent connections between these entities. Such connections can be explored using *keyword search*. Specifically, given a set of $k \geq 2$ search terms such as, e.g., "Assemblée Nationale" and "Russia", we can find all the paths that lead from a node matching the former term to another that matches the second term. This enables finding that "Assemblée Nationale" and "Russia" are connected through the contract that the wife of a member of the Assemblée has with a Russian state-owned company [2]. For a larger number of keywords, the query result will be a tree, where a keyword can be a node or an edge label in the tree.

Internship Goal: In the presence of popular search terms and/or large data graphs, there may be many answer trees connecting a given set of search terms. This raises the question of *ranking* the trees to show them to the user in the decreasing order of an *interestingness score*. A first attempt of addressing this problem is by retrieving trees with minimum weight, where the weight is the sum of edge weights. However, it is immediately clear that the sum of edge weights *fails to capture the interestingness or usefulness* of an answer. For instance, a minimum-weight answer one can find today in DBPedia, for the sample query above shows that Assemblée Nationale is an institution of France, which is a country, just like Russia. This 3-edges answer is vastly less interesting than the one exemplified above, which has 4 edges.

The intern should work toward learning what makes an answer interesting in our heterogeneous context, using human judgment. While there exists significant literature on learning to rank for information retrieval [5], they have not addressed the ranking of answer trees in a graph setting. In the first part of the project, we will gather comparative annotations of pairs of answer trees, using the labels *surprising* (does one tree bring more unexpected information than the other?) and *relevant* (is this connecting tree relevant, or representative, for the nodes it connects?) Once such valuable human input is obtained, a learning-to-rank approach can be trained to learn the structural features behind semantic notions such as relevance and surprisal. We will compare two approaches for creating feature vectors for answer trees. First, we will experiment with *hand-crafted features* such as the degree centrality of nodes in a tree [3], user profile as expressed by past queries, possibly topic modeling for different graphs and sources, e.g., [1], etc. Second, we will consider relying on *embeddings*, which are latent representations of objects and have shown remarkable results in a variety of data mining tasks [4].

Prerequisites. The intern should be familiar with machine learning concepts and have good programming skills, preferably in JAVA or Python.

Practical information. The internship will take place in the Inria team CEDAR, at Inria Saclay. A successful internship may provide opportunities for a funded Ph.D. on a follow-up subject.

References

[1] D. Card, C. Tan, and N. A. Smith. Neural Models for Documents with Metadata. *ACL*, 2018.

[2] C. Chanial, R. Dziri, H. Galhardas, J. Leblay, M.-H. Le Nguyen, and I. Manolescu. ConnectionLens: Finding Connections Across Heterogeneous Data Sources. *PVLDB*, 11:4, 2018.

[3] A. Ghazimatin, R. Saha Roy, and G. Weikum. Fairy: A framework for understanding relationships between users' actions and their social feeds. In *WSDM*, pages 240–248. ACM, 2019.

[4] W. L. Hamilton, R. Ying, and J. Leskovec. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*, 2017.

[5] T.-Y. Liu. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331, Mar. 2009.